

# MODELING DISCRETE CHOICE WITH UNCERTAIN DATA: AN AUGMENTED MNL ESTIMATOR

DANIEL HELLERSTEIN

This article introduces a multinomial logit model that uses ancillary information to control for uncertainty in both the observed choices made by respondents, and in the attributes of a respondent's choice set. Simulated data are used to compare the performance of this estimator versus simpler models, under several different kinds of uncertainty.

*Key words:* multinomial logit, simulated estimator, UIV, UDV, uncertain choices, uncertain variables.

When estimating the value of natural resources, the applied analyst must often work with noisy or otherwise imprecise measures of both dependent and independent variables. To help control for this uncertainty, this article introduces a multinomial logit model (MNL) that uses ancillary information to control for uncertainty in observed choices and in the attributes of a respondent's choice set.

For example, surveys of rural recreationists may encounter difficulties when identifying exactly where people visited—since many sites may be “informal” (such as the “the park down by the river”).<sup>1</sup> Second, even when sites can be identified, good information on site attributes may be lacking; a problem that is more likely to be true for “informal” sites (such as sites that are not intensively managed by government agencies).

When faced with such a problem, one can construct “regions,” and assume that a respondent chooses to visit a region, rather than a particular site (Feather, Hellerstein, and Hansen). Environmental attributes, such as water quality or land use, are often available on a

regional basis.<sup>2</sup> By assigning these regional measures of attributes to the constructed regions, an aggregated dataset can be created. While aggregated datasets can be used in MNL models (Ben-Akiva and Lerman), the noise introduced when regional measures are used to describe particular sites within a region can impart an “errors in variables” bias to estimated coefficients.

An additional problem arises when a significant fraction of respondents may be unsure about the actual location they visited.<sup>3</sup> In this case, the “chosen region” cannot be definitively identified. In terms of discrete choice models, this means that not only is information on the attributes of a respondent's choice set noisy, it's not always certain which alternative is chosen.

To account for the problem of uncertain data in discrete choice models (DeVaro and Lacker; Fader, Lattin, and Little), this paper introduces two augmented MNL models.<sup>4</sup> These models use Bayesian-like weighting, and

---

Daniel M. Hellerstein is a natural resource economist at the Economic Research Service of the United States Department of Agriculture. The views expressed here are those of the author, and not necessarily those of the Economic Research Service or the U.S. Department of Agriculture.

The author wish to thank Catherine Kascak and Joseph Breedlove for assisting with data collection and analysis on earlier versions of this work. Comments from an anonymous reviewer, and from the editors, are also appreciated.

<sup>1</sup> Alternatively, population surveys often are constrained in just how much detail can be acquired from respondents, either due to cost and time concerns, or due to issues of respondent confidentiality—for example, the Fishing, Hunting, and Wildlife Associated Recreation (FHWAR) survey of the U.S. Fish and Wildlife Service (<http://www.census.gov/prod/www/abs/fishing.html>).

---

<sup>2</sup> For example, the United States Department of Agriculture's Natural Resources Inventory (NRI) information that can be used to construct subcounty-level statistics on a number of land use and soil quality variables (<http://www.nhq.nrcs.usda.gov/NRI/>).

<sup>3</sup> For example, in Feather, Hellerstein and Hansen, many individuals identified a visited site using distance and direction from their home. In cases where the visited site was also identified by name, it is possible to measure the accuracy of the distance and direction information. While this distance and direction information was informative, in about one-quarter of the cases the individual incorrectly identified the basic direction (i.e., they were between 90 and 180 degrees off).

<sup>4</sup> Note that problems with uncertain data can arise in many settings other than our “rural recreation” example. For example (as suggested by a reviewer) stated preference surveys can also be subject to uncertain data; as when the analyst observes the choice made by a respondent, but with some skepticism.

simulation techniques, to control for uncertainty in both site choice and site characteristics. The next section outlines the models. Simulations are then used to test the performance of these estimators, comparing them to simpler heuristics.

### An Uncertain Dependent Variables Correction

I start with the problem of uncertainty in the dependent variable. Consider the discrete choice model, where an individual ( $i$ ) chooses from a  $k$ -element choice set ( $S$ ). Uncertainty in the dependent variable can arise when, rather than reporting an actual choice ( $S_i$ ), respondents provide uncertain information ( $D_i$ ) that the analyst then uses to imprecisely identify a chosen alternative. If, instead of simply making a best-guess as to the chosen alternative  $S_i$ , the analyst uses  $D_i$  to assign a  $K \times 1$  vector of probabilities ( $\pi_i$ ) that measure whether each alternative in  $S$  was actually chosen, then it is relatively straightforward to control for this uncertainty.

Consider our rural recreation example, where  $S$  is a set of  $k$  outdoor recreational sites, and  $D_i$  is a self-reported distance and direction (from the respondent's home to where she recreated). Since people may have a hazy idea as to just where they went, the reported distances and directions may be imprecise. However, if a sub-sample of visitors provide the actual site visited (say, a site name), along with distance and direction, then a reporting-accuracy model can be constructed by regressing the actual distance and direction (to the visited site) against the self reported distance and direction. Using such a reporting-accuracy model, the goal is to generate (for each respondent) the probabilities ( $\pi_1, \dots, \pi_k$ ) of actually visiting each of the available sites ( $S_1, \dots, S_k$ ).<sup>5</sup>

Incorporating  $\pi$  into an MNL model yields the MNL model with an Uncertain Dependent Variables (UDV) correction. This yields a Bayesian-like expression that computes a predicted probability of choosing an alternative by combining prior probabilities with the observed attributes of these alternatives. For a

single respondent  $i$ , who reports  $D_i$ , the contribution to the likelihood function is

$$(1) \quad \mathcal{L}_i = \sum_{k=1}^K \left( \frac{\exp(\mu_{ik})}{\sum_{\kappa=1}^K \exp(\mu_{i\kappa})} \pi_{ik} \right) \\ = \left( \frac{\sum_{k=1}^K \pi_{ik} \exp(\mu_{ik})}{\sum_{\kappa=1}^K \exp(\mu_{i\kappa})} \right)$$

where  $K$  is the total alternatives to choose from,  $\mu_{ik} = X_{ik}\beta$  is the  $i$ th respondent's systematic utility from choosing alternative  $k$  (of  $1, \dots, K$  alternatives); given observable alternative-specific attributes  $X_{ik}$ . Actual utility equals:  $\mu_{ik} + \epsilon_{ik}$  ( $\epsilon_{ik}$  an unobservable random factor with an extreme value distribution). The UDV correction  $\pi_{ik} = \pi(k|D_i)$ : the "prior" probability that respondent  $i$  actually chose alternative  $k$ , given that she reported  $D_i$ . Please see the Appendix for a discussion of how these can be derived using a reporting-accuracy model in conjunction with a Bayesian mechanism. The log likelihood for the entire sample ( $i = 1, \dots, I$ ) of respondents is

$$(2) \quad \ln \mathcal{L} = \ln \left( \prod_{i=1}^I \mathcal{L}_i \right) \\ = \sum_i \left[ \ln \left( \sum_{k=1}^K \pi_{ik} \exp(X_{ik}\beta) \right) - \ln \left( \sum_{k=1}^K \exp(X_{ik}\beta) \right) \right].$$

Note that when the choice of  $k'$  is known with certainty ( $\pi_{k'} = 1$ ;  $k = 1, \dots, k' - 1, k' + 1, \dots, K$ ; and  $\pi_k = 0$ ), equation (2) collapses to the standard MNL. The gradient of equation (2), which is described in the Appendix, is also similar to the gradient of the standard MNL.

### An Uncertain Independent Variables Correction

In the MNL model with a UDV correction discussed above, the dependent variable is not known with certainty, but the independent variables are precisely measured. That is, although the analyst may not be certain which alternative was chosen, she does have accurate measures of the attributes of these alternatives. However, such accurate information may not be available, forcing the analyst to use

<sup>5</sup> For respondents reporting an actual site name, the probability for the actual site will equal 1, with all other probabilities set equal to 0.

noisy measures, and thereby risking an errors-in-variables bias.

For example, in our rural recreation model the obscurity of many recreational sites may force the analyst to use regional averages, such as county averages of land use and water quality, as measures of site attributes. However, since regions are usually not homogeneous, using regional averages introduces a possibility of errors in variables.<sup>6</sup>

In order to minimize this type of bias, we consider an Uncertain Independent Variables (UIV) correction, where noisy measures of independent variables are explicitly controlled for.

The model requires an additional set of information: the variance/covariance matrix of the independent variables. In general, this is required on a per region basis; though in some special cases a single matrix could describe all regions (if regions differ in average measures, but not in higher moments).<sup>7</sup>

Given this information, the likelihood for observation  $i$  is

$$(3) \quad \mathcal{L} = \int_{\chi} \left[ \sum_{k=1}^K \left( \frac{\exp(X_{ik}\beta)}{\sum_{\kappa=1}^K \exp(X_{i\kappa}\beta)} \pi_{ik} \right) \right] f(X) dX$$

where  $\chi$  is range of support of  $X$ ,  $X$  is  $X_{i1} \dots X_{iK}$ ,  $X_{ik}$  is  $J \times 1$  vector of explanatory variables for choice  $k$ , and  $f(X)$  is probability of observing  $X$ .

Note the inclusion of the  $\pi_{ik}$  UDV correction in equation (3). If the chosen site is known with certainty,  $\pi_{ik}$  can be set to 1.0 (for the chosen site) and 0.0 (for all other sites).

Also note that this requires integrating a  $J^*K$  multivariate distribution (say, a multivariate normal) over the (possibly  $-\infty$  to  $+\infty$ ) range of support of  $X$ . To reduce this complexity one can use

$$\begin{aligned} \mu_{ik} &= X_{ik}\beta \\ E[\mu_{ik}] &= \bar{X}_{ik}\beta \\ \text{Var}[\mu_{ik}] &= \beta' \Sigma_{ik} \beta \end{aligned}$$

where  $\bar{X}_{ik}$  is the expected value of  $X_{ik}$ , and  $\Sigma_{ik}$  is variance matrix of  $X_{ik}$ .

<sup>6</sup> Alternatively, consider a model of recreational fishing, where the catch rate at sites depends on unobservable angler-specific attributes. Using site-specific average catch rates to explain an individual's site choice would introduce errors in variables problem.

<sup>7</sup> Note that "region" should be read as a shorthand for "the set of imprecise measures, often derived from aggregate data, used to assign values to the attributes of an alternative."

Substituting  $\mu$  for  $X\beta$

$$(4) \quad \mathcal{L}_i = \int_{\bar{M}} \left[ \sum_{k=1}^K \left( \frac{\exp(\mu_{ik})}{\sum_{\kappa=1}^K \exp(\mu_{i\kappa})} \pi_{ik} \right) \right] \times f(U) dU$$

or

$$\ln \mathcal{L}_i = \ln \left( \int_{\bar{M}} \left[ \sum_{k=1}^K \left( \frac{\exp(\mu_{ik})}{\sum_{\kappa=1}^K \exp(\mu_{i\kappa})} \pi_{ik} \right) \right] \times f(U) dU \right)$$

where  $\bar{M}$  is range of support of  $U$ ,  $U$  is  $\mu_{i1} \dots \mu_{iK}$ , and  $f(U)$  is probability of observing  $U$ .

It is often reasonable to assume that the realization of  $X_{ik}$  is independent of  $X_{il}$  ( $l \neq k$ ). Hence, the variance matrix of  $X$  is block diagonal and the variance matrix of  $U$  is diagonal.

Given the complexities of multivariate integration over a normal (or other) density function, simulation methods can be used to solve the above. In particular, for each observation a total of  $R$  different  $K \times 1$  vectors ( $M_r$ ,  $r = 1, \dots, R$ ) of critical values are drawn from a multivariate normal distribution. Hence (suppressing the  $i$  subscript):

$$M_r \sim N(\bar{M}, \Omega)$$

where  $\bar{M} = E[u_1], \dots, E[u_K]$

$$(5) \quad \Omega = \begin{pmatrix} \omega_1 & \dots & 0 \\ \dots & \omega_2 & \dots \\ \dots & \dots & \omega_K \end{pmatrix}$$

where  $E[\mu_k] = \bar{X}_k \beta$ ,  $\bar{X}_k$  is expected value of  $X_k$ ,  $\omega_k$  is  $\text{Var}[\mu_k] = \beta' \Sigma_k \beta$ , and  $\Sigma_k$  is variance matrix of  $X_k$ .

Following the method discussed in Train, for each  $M_r$  ( $K \times 1$ ) vector of values, the integrand of equation (4) is calculated. The likelihood, for a given observation and a particular value of  $\beta$ , is the average of the  $R$  different values—with one value for each realization of  $M_r$  ( $r = 1, \dots, R$ ).

Note that  $\Sigma_k$  is the covariance of the independent variables of individual  $i$ 's  $k$ th choice, and  $\omega_k$  is the variance of  $\mu_k$ . Also, due to inter-regional independence of  $X$ ,  $\Omega$  is a  $K \times K$  diagonal matrix.

Re-expressing equation (4) (and noting that equation (4) incorporates the UDV

correction) yields the Uncertain Independent and Dependent Variables (UIDV) model:

$$(6) \quad \ln \mathcal{L}_i = \ln \left( \sum_{r=1}^R \frac{\left[ \frac{\sum_{k=1}^K \exp(\mu_{ikr}) \pi_{ik}}{\sum_{k=1}^K \exp(\mu_{ikr})} \right]}{R} \right)$$

where  $\mu_{ikr}$  is  $\bar{X}_{ik}\beta + \delta_r \sqrt{\beta' \sum_{ik} \beta}$  and  $\delta_r$  is a draw from a standard normal.

Note that the second term in  $\mu_{ikr}$  captures the effect of uncertainty-in-X on the value of  $\mu$ . For the gradient of equation (6), please see the Appendix.

### Simulated Data

To investigate the performance of the UIV and UDV corrections, we turn to a simulation.<sup>8</sup> The simulation has several components: generation of a set of sites and a “population of visitors,” computation of which site is selected and which site appears to be selected, estimation of coefficients using several models, computation of welfare measures, and comparison of results.

#### Site Generation

The basic simulated landscape consists of a grid of cells, with each cell tantamount to a “region.” Each of these regions is described by a value for each of several “regional” attributes. The spatial pattern of these “regional” attributes can vary from totally uniform (same value in each cell), to random, to parametrically described.

Sites are then randomly scattered across the grid. Depending on the region (the cell) that a site lands in, values for each of several attributes are assigned to the site. This assignment uses the “regional” attributes as means of a uniform distribution. Thus, several sites within a single region will have similar, but not exactly the same, attributes.

#### Population Generation

Individuals are randomly scattered across the grid.

### Actual Site Selection

For each individual ( $i$ ), the utility offered by each of the sites is computed as

$$U_{is} = S_s * \beta + D_{is} * \beta_d + \epsilon_{is}$$

where  $S_s$  are the attributes of site  $s$ ,  $D_{is}$  is the distance from individual  $i$  to site  $s$ ,  $\epsilon_{is}$  is the extreme value distributed random error, and  $\beta_s, \beta_d$  are the coefficients. Note that  $\beta_d < 0$ .

The chosen site is the one with largest  $U_{is}$ .

### Reported Site Selection

Two methods are used to simulate the effects of using less than perfect site selection information:

1. Using the “region” the site is in. That is, rather than using the attributes of the site, use the “regional” attributes of the cell that the site is in. This simulates a case where attributes of sites are not known, but statistics on environmental attributes are known at a regional level.
2. A “reported distance/direction” reported site is generated by randomly deviating both the distance and direction traveled by the individual. For sufficiently large deviations, the location pointed to by the reported distance/direction will be in a different region than the actually visited site. This simulates uncertain information about which site was chosen, even when aggregations of site (“regions”) are used as choices.

### Model Estimation

An MNL model is used to estimate coefficient vector for the true model (the model using the actual site choice). MNL models, including the UDV and UIV corrections, are also used to estimate a coefficient vector for each of the less-than-perfect site choice models.

A simple heuristic is used for models incorporating the UDV and UIV corrections—the probability of reporting a distance/direction given that a particular site was actually chosen,  $p(D_i|k)$ , is inversely proportional to the distance between the “reported site” and the “actual site.”

### Welfare Estimation

For each estimated coefficient vector a compensating variation for each site choice

<sup>8</sup> The GAUSS program used to create the simulated datasets, and to estimate the models, is available from the author upon request.

occasion computed as (Ben-Akiva and Lerman):

(7)

$$CV = \ln \frac{\left(\sum_{s=1}^S \exp(S_s \beta + D_{is} \beta_d) + 0.577\right)}{\beta_d}.$$

Comparison of Results

For each individual, the CV computed using coefficients generated by a less-than-perfect model is compared to the true CV. The true CV is calculated using the attributes of the site actually chosen, and the true value of  $\beta$ .

Our first results, in table 1, display how effective several heuristics are in dealing with uncertainty. Each row represents a single simulation, with each column a different MNL estimator. To reiterate, none of the models in table 1 use the UIV or UDV corrections.

These results (which are representative of numerous runs using different deviations, different random number seeds, and different values of the true beta) reveal few clear patterns. In general, the “actual site, estimated beta” (column 1) model performs well. The “actual region, actual beta” (column 2) model has errors that never exceed 30%, which sug-

gest that the use of aggregate sites does not introduce overwhelming errors. However, use of estimated beta with the “actual region” (column 3) has mediocre performance. In fact, all models that use regional data and estimated betas perform roughly equally well. The “actual” region results tend to be best, but not always.

The “50% correct region & 50% reported region” (column 5) model reflects a “use all available information” scenario: the notion is that one-half of respondents do not go to identifiable sites, forcing the analyst to use distance and direction information. These results are somewhat better than the “reported” (column 4) results, possibly reflecting less error in the dependent variable. However, the results are not across the board (sometimes the “reported” results are better).

With this as background, consider MNL models that contain the UDV and UIV corrections. Table 2 displays the results from a collection of simulations that incorporate UDV and UIV corrections. Each cell represents averages over a number of simulations using the same coefficient and model structure.

Summarizing (and these results are representative of a number of simulations), the last two rows indicate that, by itself, the use of the UDV correction does not seem to help.

Table 1. Selected Simulation Results. Percentage Deviation of Estimated CV from True CV

	Actual Site, Est. $\beta$	Actual Region, Actual $\beta$	Actual Region Est. $\beta$	Reported Region, Est. $\beta$	50% Actual Region & 50% Reported Region, Est. $\beta$
<i>Constant “regional averages”</i>					
Low error	26	24	205	133	162
Medium error	24	20	104	56	46
High error	3	27	147	75	32
<i>Random “regional averages”</i>					
Low error	5	21	77	75	77
Medium error	20	21	64	86	78
High error	3	24	15	28	13
<i>Normal “regional averages”</i>					
Low error	9	19	25	17	11
Medium error	5	18	76	67	71
High error	22	21	14	44	17

Notes: The true average CV ranges between 5 and 10. Smaller values indicate better performing models (small deviations from the “true” value).  
The low, medium, and high error rows refer to the size of the error committed by respondents when reporting their site choice. Inaccuracy is driven by deviating the reported distance and direction using a uniform random number with a limits of (plus or minus):

	Distance	Direction
Low	10%	10
Medium	25%	40
High	40%	60

For distance, the limits are relative to the actual distance. For direction, the standard deviation is in angular degrees.  
The “constant,” “random,” and “normal” sub-table labels refer to how regional average values (for site characteristics) were assigned. “Constant” means that all regions have the same average value, “random” means that each region is randomly assigned an average value, and “normal” means that each region’s value is derived from a two-dimensional normal distribution.

**Table 2. Selected Simulation Results. Average Percentage Deviation of Estimated CV from True CV**

Model	Without UIV Correction	With UIV Correction
Actual site	39 (68)	n.a.
Actual region	42 (23)	32 (19)
Reported region	65 (36)	49 (24)
50% actual region & 50% reported region	47 (21)	32 (19)
Reported region, with UDV	73 (47)	31 (17)
50% actual region & 50% reported region, with UDV	48 (23)	32 (21)

Notes: Standard deviations are in parenthesis. The average % deviation when using the actual region, and actual  $\beta$  is 15% (with a 3% standard deviation). For the actual site mode, the UIV is not applicable, since attributes are known with certainty. Averages, and standard deviations, are over thirty-six simulations, using twelve different scenarios. Similar results were obtained when ninety-six simulations were performed (on twelve scenarios). The twelve scenarios differ in the reporting accuracy (the size of the distance/direction errors), and how regional averages are distributed (uniform, random, or normal). The "error" in predicting the compensating variation is computed as:  $100 \times (\text{predicted\_cv} - \text{actual\_cv}) / \text{actual\_cv}$ . "UDV" refers to the uncertain-dependent variables correction. In the "50% actual region & 50% reported region, with UDV" model, one half the observations utilize the UDV correction. The other half use the actual region (observations have  $\pi_k = 1$ , all other  $\pi_k = 0$ , for some site  $k$ ).

In fact, it can be worse than using the distance/direction models without a UDV correction (row 4 vs. row 6), though in most simulations it was about the same. Moreover, comparing row 4 to row 3, and row 6 to row 5, shows that the use of actual destination data (if available) is almost always a good idea.

However, the disappointing results from using the UDV correction are largely resolved (with errors cut in half) when the UIV correction is also applied. Not surprisingly, the UIV correction also improves the actual region model (row 2), since it helps control for possible errors in variables due to the use of regional aggregation. Furthermore, the UIV correction also improves the simple distance/direction models (row 5 vs. row 3).

It would seem, for data that contain uncertain measures of both independent and dependent variables, that correcting for variance in the independent variable (the UIV correction) is more important than a correcting for variance in the dependent variable (the UDV correction), though doing both is best.

## Conclusions

When considering valuation problems where information on site choice is uncertain, the

analyst can turn to alternative sources of data. These include alternative measures of "what site was chosen," such as the use of distance and direction information; and it includes alternative measures of the attributes of each choice in a respondent's choice set, as can be provided by using regional aggregate measures instead of unobservable site-specific attributes.

In this paper, we introduce two corrections to the MNL estimator to deal with these sources of uncertainty. The Uncertain Dependent Variables (UDV) correction uses a Bayesian-like weighting, which assigns to each alternative a probability that it was chosen, a probability based on imprecise information as to what the respondent actually chose. The Uncertain Independent Variables (UIV) correction is an errors-in-variable type correction, incorporating an iterative algorithm that uses the variance of regional attributes as a means to control for aggregation bias.

Using simulated datasets, several different models (that contain imprecise measures of the dependent and independent variables) were examined. These include models that incorporate the UDV and UIV corrections. By itself, the performance of the UDV correction is disappointing. However, when combined with the UIV correction, substantial improvements occur, with aggregate models performing almost as well as models that incorporate exact information.

These results suggest that analysts faced with the problem of valuing "obscure" sites (such as may be common in rural areas), and sites for which noisy data on independent variables is available (such as regional biophysical measures), ought to consider use of models that explicitly incorporate uncertainty in site choice and site attributes.

I conclude by noting that the simulations examined here are far from complete. Further analysis, focusing on simpler cases (such as when site choice is uncertain, but site attributes are known) is called for. Additional measurement issues include the effects of imprecision in the reporting-accuracy model, and the impacts of approximating the covariance of independent variables. A more complicated, iterative model may also merit attention: which updates non-diffuse priors, that are then combined with reporting probabilities to compute Bayesian calculations of prior probabilities.

[Received September 2002;  
accepted May 2004.]

## References

- Ben-Akiva, M., and S. Lerman. *Discrete Choice Analysis: Theory and Application to Travel Demand*. Cambridge MA: The MIT Press, 1985.
- DeVaro, J., and J. Lacker. "Errors in Variables and Lending Discrimination." *Economic Quarterly* 81(1995):19–31.
- Fader, P., J. Lattin, and J. Little. "Estimating Non-linear Parameters in the Multinomial Logit Model." *Marketing Science* 11(1992):372–85.
- Feather, P., D. Hellerstein, and L. Hansen. "Economic Valuation of Environmental Benefits of the Targeting of Conservation Programs." *Agricultural Economics Report Number 778*. Washington DC: Economic Research Service, United States Department of Agriculture, 1999.
- Train, K. "Recreation Demand Models with Taste Differences Over People." *Land Economics* 74(1998):230–39.

## Appendix

### Computing Prior Probabilities

In many cases, the prior probability used in the UDV model ( $\pi$ ) may be determined using a Bayesian method, where the analysts starts with the "reporting" probability:

$$(A.1) \quad \pi_{ik} = \frac{p(D_i|k)}{\sum_{\kappa=1}^K p(D_i|\kappa)}$$

where  $p(D_i|k)$  is the "reporting probability" of respondent  $i$ .

The reporting probability is a density function that describes the probability of reporting  $D_i$  given that alternative  $k$  was actually chosen. For example, by centering a two-dimensional normal distribution on the location of the actual site chosen ( $L_{ik}$ ),  $p(D_i|k)$  would be the probability value read at point  $D_i$  (suitably offset from  $L_{ik}$ ).

A more complete model could incorporate a number of modifications. For example, individual attributes ( $Z$ ) might determine respondent accuracy, yielding  $p(k|D_i, Z_i)$ . Or, some site-specific attributes ( $L$ ) may be correlated with how easy it is to report their location, yielding  $p(k|D_i, L_k)$ .

Finally, the reporting-probability model need not start with the diffuse priors of each site being chosen with equal probability. Instead, prior probabilities could be incorporated:

$$(A.1a) \quad \pi_{ik} = \frac{\tau_k p(D_i|k)}{\sum_{\kappa=1}^K \tau_k p(D_i|\kappa)}$$

where  $\tau_k$  is a prior probability of choosing alternative  $k$ . In (A.1),  $\tau_k = 1$  for all  $k$ . A non-diffuse prior would modify  $\tau_k$  using other information; for example, information on the total number of people choosing each alternative. Alternatively, an iterative strategy could be used, with  $\tau_k = 1$  in the first round, and in latter rounds  $\tau_k$  is set equal to the predicted probability of choosing alternative  $k$ .

Please note that the "prior" probability  $\tau_k$  is not the same as the "prior" probability  $\beta_{ik}$ .  $\pi$  is an exogenously determined probability of choosing alternative  $k$ ; in the simplest case, it is 1.0 for all alternatives.  $\pi$  is the probability of choosing alternative  $k$ , given respondent-specific information ( $D$ ), a reporting probability model ( $p$ ), and  $\tau$ . In the simplest case, it is 1.0 for one alternative and 0.0 for all other alternatives,

### Gradient of the UDV Model

The gradient of the log likelihood of the UDV model is similar to the standard MNL:

$$(A.2) \quad d \ln \mathcal{L} d\beta = \sum_i \left[ \frac{\sum_k \pi_{ik} X_{ik} \exp(X_{ik}\beta)}{\sum_k \pi_{ik} \exp(X_{ik}\beta)} - \frac{\sum_k X_{ik} \exp(X_{ik}\beta)}{\sum_k \exp(X_{ik}\beta)} \right].$$

### Gradient of the UDV/UIV Model

The gradient of the model with both UIV and UDV corrections (equation [6]), with respect to  $\beta$ , can also be approximated using equation (4) and the critical values from equation (5). We use the chain rule

$$(A.3a) \quad \frac{dG}{dx} = \frac{d(g(f_1(x), f_2(x), \dots, f_K(x)))}{dx} = \frac{dg}{df_1} \frac{df_1}{dx} + \frac{dg}{df_2} \frac{df_2}{dx} + \dots + \frac{dg}{df_K} \frac{df_K}{dx}.$$

The "dg" terms in equation (A.3a) are

$$(A.3b) \quad d \ln \mathcal{L}_i / d\mu_{ikr} = \frac{(\exp(\mu_{ikr})\pi_{ik} \times \sum_{\kappa} \exp(\mu_{ikr})) - (\exp(\mu_{ikr}) \times \sum_{\kappa} \exp(\mu_{ikr})\pi_{ik})}{(\sum_{\kappa} \exp(\mu_{ikr}))^2}.$$

The “ $df$ ” terms are

$$(A.3c) \quad \frac{d\mu_{irk}}{d\beta} = X_{ik} + \delta_r \frac{\sum_{ik} \beta}{\sqrt{\beta \sum_{ik} \beta}}.$$

Note that the  $R$  divisor terms cancel out, so they do not appear in equation (A.3d).

---

Substituting, we get

(A.3d)

$d \ln \mathcal{L}_i / d\beta$

$$= \frac{\sum_{r=1}^R \left( \sum_{k=1}^K \left[ \frac{(\exp(\mu_{ikr}) \pi_{ik} \times \sum_{\kappa} \exp(\mu_{ikr})) - (\exp(\mu_{ikr}) \times \sum_{\kappa} \exp(\mu_{ikr}) \pi_{ik})}{(\sum_{\kappa} \exp(\mu_{ikr}))^2} \times \left( X_{ik} + \delta_r \frac{\sum_{ik} \beta}{\sqrt{\beta \sum_{ik} \beta}} \right) \right] \right)}{\sum_{r=1}^R \left[ \frac{\sum_{k=1}^K \exp(\mu_{ikr}) \pi_{ik}}{\sum_{k=1}^K \exp(\mu_{ikr})} \right]}.$$